

# Towards meeting the central challenge in robot ethics

**Bentzen, Martin Mose**

*Publication date:*
2016

*Document Version*
Peer reviewed version

**Towards meeting the central challenge in robot ethics**

The development of a robot ethics, which can be implemented in robots to ensure the safety of their actions from an ethical point of view will require insights from both engineering and ethics, see e.g. (Wallach and Allen, 2008), (Arkin, 2009), (Lina, Abney, and Bekey (eds.), 2012), (Winfield et alt., 2014) . On the one hand, an implementation of the robot ethics needs to be technically feasible. On the other hand, the principles upon which the robot ethics rests must be justifiable and clear from an ethical perspective. However, the gap between robot engineering and informal ethics is big. We argue that there are strong reasons for seeing bridging this gap as *the central challenge in robot ethics* today. Reflections on the central challenge delimit the scope of the current work. Of course this challenge can be met in various ways. Further, the aim of this work is not to solve all problems involved with meeting the central challenge, but rather to provide stepping stones in the right direction. As our *basic methodological claim,* we argue that tools from formal modelling, e.g. formal semantics and formal logic, can provide ways to build bridges between the seemingly disparate areas: engineering and informal ethics. As formal semantics and logic are mathematically precise, a formal theory is an important step towards reliable implementation. As the tools are conceptually rigorous and clear they provide a way of explicating philosophical intuitions, making assumptions clear, and forcing conceptual choices to be made, e.g. about what logical properties certain concepts have. See (Bringsjord, S., Arkoudas, K., Bello, P. 2006) for a seminal statement and actual implementation of a methodology similar to the one advocated here.

On the more technical side, in this paper, tools known from the logic of agency are combined with tools from the logic of causality to model situations with ethical robots. This gives an expressive abstract framework, in which we can model intentions, causal effects and values of actions of various agents. More specifically, the models are formally related to multi-agent single moment models known from STIT theory, see e.g. (Horty, 2001), combined with directed acyclic graphs (DAGs) known from Judea Pearl's structural approach to causality, see (Pearl, 2009). This makes it possible to formalize basic causal notions, such as various kinds of causal influence and causal overdetermination of an event. Consequences of actions are ordered as a causal network with actions of agents at the root of the network, i.e. action tokens are background variables in Pearl's terminology. When the action of each agent is decided, the causal consequences of the variables relevant to the situation can be calculated. The actions and consequences of agents thus form a deterministic causal mechanism, again in Pearl's terminology. There is however, epistemic or ontological indeterminism as to the values of the background variables, i.e. which action is performed by each agent. The framework is neutral with regard to whether this indeterminism is actually epistemic or ontological, and it is not ruled out that factors outside the framework decide which action agents actually perform, as will be the case for robots. The causal network approach makes it possible to e.g. distinguish agents with causal influence on a particular consequence from bystanders without causal influence. Intentions are modelled as sets of consequences. Including intentions makes it possible to formalize e.g. unintended side-effects of actions. The framework can be used prospectively, simulating the possible outcomes of different combinations of actions, and retrospectively, evaluating causal responsibility for different effects of actions. The framework also makes it possible to model ethical principles that go beyond consequentialism (thus a step forward from e.g. (Horty, 2001)). In particular we show how to model a causal version of pareto-optimization, and the double effect principle, see e.g. (Mangan, 1949), (Foot, 1967), (McIntyre, 2014). The intuition behind the causal version of pareto optimization is that the agent should act such that the causal effects of its action makes a situation better for most involved agents

without making it worse for one single agent. We argue that pareto-optimization is a fitting principle for robots acting in most civilian contexts, e.g. for autonomous vehicles and many rescue robots. However, for robots in disaster situations where many lives are potentially at stake, and in military contexts, the robot may be required to act according to less strict principles such as the double effect principle or even (in very remote cases) simple utilitarianism. We thus argue for a *principled ethical contextualism* for ethical robots, whose basic idea is that various principles are complementary and apply in different contexts. We consider limitations and possible objections to this position and e.g. discuss the difficulties involved in finding precise criteria for differentiating contexts where various principles apply.


Arkin R (2009). *Governing Lethal Behavior in Autonomous Robots*. CRC Press

Bringsjord, S., Arkoudas, K., Bello, P. 2006. Toward a general logicist methodology for engineering ethically correct robots, *IEEE Intelligent Systems*, 21 (4): 38–44.

Foot P (1967). "The problem of abortion and the doctrine of double effect", *Oxford Review* 5:5–15

Horty JF (2001) *Agency and Deontic Logic*. Oxford University Press

Lina, P., Abney, K., Bekey, G.A. (eds.) 2012. *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press.

Mangan J (1949). "An historical analysis of the principle of double effect", *Theological Studies* 10:41–61

McIntyre A (2014) "Doctrine of double effect", in: Zalta EN (ed) *The Stanford Encyclopedia of Philosophy*, winter 2014 edn

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference,* (2nd edn). Cambridge University Press.

Wallach W, Allen C (2008) Moral Machines: Teaching Robots Right From Wrong. Oxford University Press

Winfield AF, Blum C, Liu W (2014) "Towards an ethical robot: internal models, consequences and ethical action selection", in: Mistry M, Leonardis A,MWitkowski, Melhuish C (eds) Advances in Autonomous Robotics Systems, Springer, pp 85–96.